Ulrike Kuhl
Bielefeld University, Bielefeld, Germany

Mark T. Keane
University College Dublin, Dublin, Ireland

Plan for the Day

CF Tutorial

TIME	Topics
9:00 AM	***Introduction***
	Hello and Introducing Ourselves!
	Hands-on: *Trying Our Study (follow link)*
9:30 AM	***Historical Fundamentals of Counterfactuals***
	From Philosophy to XAI (via Psychology)
	Two Sample User Studies and Q&A
10:30 AM	COFFEE (10:30-11:00)
11:00 AM	***Fundamentals of Counterfactuals in AI***
	Formalisation
	Modelling Approaches & Key Constraints
11:30 AM	***Using Counterfactual Algorithms***
	Hands-on: *A Counterfactual Toolbox (AA)*
	Hands-on: *Checking Out Notebooks and Q&A*
12:00 PM	***Fundamentals of User Studies***
	User Studies I: A Simple Two-Group Design
12:30 PM	LUNCH (12:30-14:00)
2:00 PM	***Algorithmic Growth Points***
	Computational Future Directions and Q&A
2:30 PM	***More Fundamentals of User Studies***
	User Studies I: A Simple Two-Group Design (cont.)
3:00 PM	COFFEE (15:00-15:30)
3:30 PM	***From Fundamentals to an Actual User Study***
	User Studies II: A More Complex Design
	User Studies III: Even More Complex Designs
	Hands-on: *Looking At Our Study*
5:00 PM	***Closing Session, Discussion and Final Q&A***
	TUTORIAL END

Six Steps to Heaven…

Motivation

Design

Materials & Procedure

Piloting

Data Collection & Analysis

Results

Motivation…Prior Stuff

- To test for *Effects of Explanation Use*
- Consider Ethical Issues (permission?)
- Pre-register the Study (more common)
- Bias Minimizing:

  ~ Researcher bias (are you biased)
  ~ Avoid testing null hypothesis
  ~ Participant biases (selection, social desirable)

Let's Crawl Before Running…

Design

- Beyond the single group study (rare), the two-group study is a go-to standard

- To test for effects of explanation use

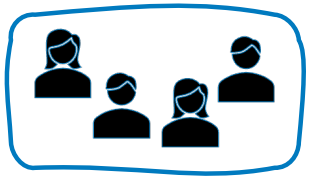
Between-Subjects

--> 2 groups in 2 conditions

Experimental Group:
CF Explanation (CF-XP)

vs.

Control Group:
No Explanation (No-XP)



Two Groups

Control:
No-Explanation
(NO-XP)

Experimental:
CF-Explanation
(CF-XP)

What are we going to show people?

STEP 2

Design

Both See Same Items

Materials & Procedure

Control: No-Explanation (NO-XP)

Experimental: CF-Explanation (CF-XP)

What are we going to show people?

Both See Same Items

Materials & Procedure

data-set/domain

random sample (40 from 1000s)

James	
Gender	Male
Weight	81kg
Units	6
Duration	105 mins
Stomach	Full
Limit	**?**

material-set (both variants)

What are we going to show people?

Single Group Study

Materials & Procedure

constrained sample

data-set/domain

James	
Gender	Male
Weight	81kg
Units	6
Duration	105 mins
Stomach	Full
Limit	?

- ***with some (not all) features***
- ***near decision boundary***
- ***outcomes > 0.7 certainty***
- ***all with positive outcomes***
- ***all with negative outcomes***

…and so on

material-set

Do items have key features we want?

Is Material-Set Balanced?

STEP 3

Materials & Procedure

item1	item2	item3	item4	item5	item6	item7	item8	item9	Item10

½ items are positive ½ items are negative

Is the material set balanced?

Is Material-Set Balanced?

STEP 3

Materials & Procedure

item1	item2	item3	item4	item5	item6	item7	item8	item9	Item10

½ items are positive ½ items are negative

item5	item1	item6	item3	item10	item2	item9	item4	item7	Item8

item10	item2	item6	item8	item5	item4	item7	item1	item9	Item3

⋮

But, randomly re-ordered items for each person !

Is Material-Set Balanced?

item1	item2	item3	item4	item5	item6	item7	item8	item9	Item10
Feat-A	Feat-B	Feat-C	Feat-D	Feat-E	Feat-A	Feat-B	Feat-C	Feat-D	Feat-E

½ items are positive ½ items are negative

each item uses one of five key features

Is the material set balanced for key variables?

Is Material-Set Balanced?

item1	item2	item3	item4	item5	item6	item7	item8	item9	Item10
M **Feat-A**	F **Feat-B**	M **Feat-C**	F **Feat-D**	M **Feat-E**	F **Feat-A**	M **Feat-B**	F **Feat-C**	M **Feat-D**	F **Feat-E**

½ items are positive ½ items are negative

each item uses one of five key features

½ items have male gendered, ½ items have female gendered

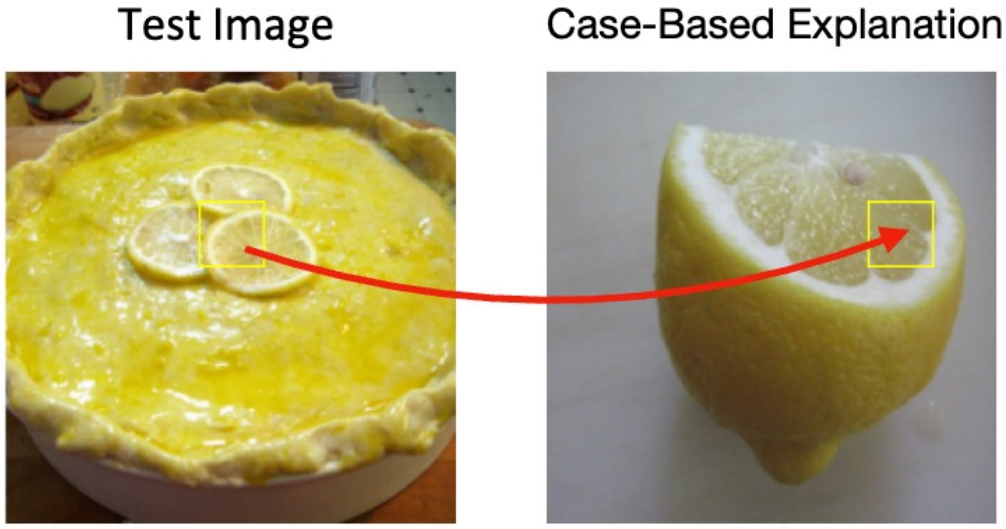
Is the material set balanced for **hidden** variables?

Is Material-Set Balanced?

This Can Get Messy…

Materials & Procedure

- Even a randomly chosen material-set can end up eliciting different responses (unpredicted)
- Sometimes balancing is really hard or impossible?
- Hidden variables are just intuitive, eyeball?
- >1 group ups ante !



What Do They Do?

STEP 3

Materials & Procedure

Control: No-Explanation (NO-XP)

Experimental: CF-Explanation (CF-XP)

James	
Gender	Male
Weight	81kg
Units	6
Duration	105 mins
Stomach	Full
Limit	**?**

NO-XP

Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
(1)	(2)	(3)	(4)	(5)

NEXT ITEM

CF-XP

If James had drunk 5 units, he would be under the limit.

Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
(1)	(2)	(3)	(4)	(5)

Read the item, make a prediction, give confidence rating, get No/CF-XP…(very simple, clear insts)

What Do They Do?

Control: No-Explanation (NO-XP)

Experimental: CF-Explanation (CF-XP)

James	
Gender	Male
Weight	81kg
Units	6
Duration	105 mins
Stomach	Full
Limit	**?**

STEP 3

Materials & Procedure

NO-XP

Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
(1)	(2)	(3)	(4)	(5)

NEXT ITEM

CF-XP

If James had drunk 5 units, he would be under the limit.

Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
(1)	(2)	(3)	(4)	(5)

Can't Be the Same ? But Need to be Matched!

Cognitive load and time on task needs to be similar.

What Do They Do?

Materials & Procedure

Control: No-Explanation (NO-XP)

Experimental: CF-Explanation (CF-XP)

James	
Gender	Male
Weight	81kg
Units	6
Duration	105 mins
Stomach	Full
Limit	**?**

NO-XP

James is over the limit, because of his features.

Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
(1)	(2)	(3)	(4)	(5)

CF-XP

If James had drunk 5 units, he would be under the limit.

Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
(1)	(2)	(3)	(4)	(5)

Can't Be the Same ? But Need to be Matched!

Cognitive load and time on task needs to be similar.

The Hidden Stuff

That Can Go Un-noticed

Materials & Procedure

- **Active vs. passive**: Asking people to produce a prediction → engagement
- Correct/Incorrect answer is an ***objective measure***
- Care with ***subjective measures***, when and what; can they really make that judgement reliably (confidence now or later); correct vs. reasonable?
- People take short-cuts, inattention (checks), responding to some extraneous factors (T, T -> T)

The Non-Hidden Stuff

Boilerplate: Consent, Permissions & Contact Information

Materials & Procedure

- Informed Consent
  - ~ Clear detail of study aim & procedure
  - ~ Voluntariness, right to withdraw
  - ~ Reimbursement & contact person

- Privacy and Confidentiality
  - ~ Statement of confidentiality
  - ~ Avoid personal data
  - ~ Properly anonymized

A Word On: Between V Within

Within Can Be Better, If Possible

Materials & Procedure

Control: No-Explanation (NO-XP)

Experimental: CF-Explanation (CF-XP)

Between-Participants Design

Control: No-Explanation (NO-XP)

Experimental: CF-Explanation (CF-XP)

Within-Participants Design

Maybe less variance: people are their own controls

(but: beware of fatigue, randomise order of conditions, …)

Formats for Data Acquisition

Data Collection & Analysis

Online

Laboratory

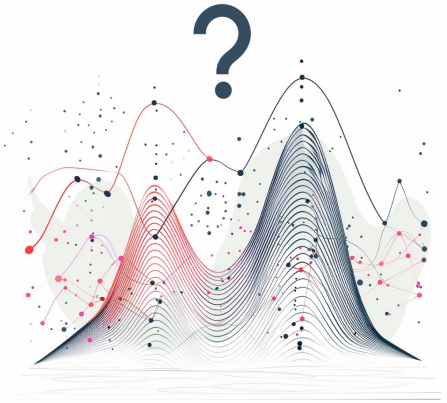
Online will tend to be cruder…

Piloting

Pilot: Why Do Them?

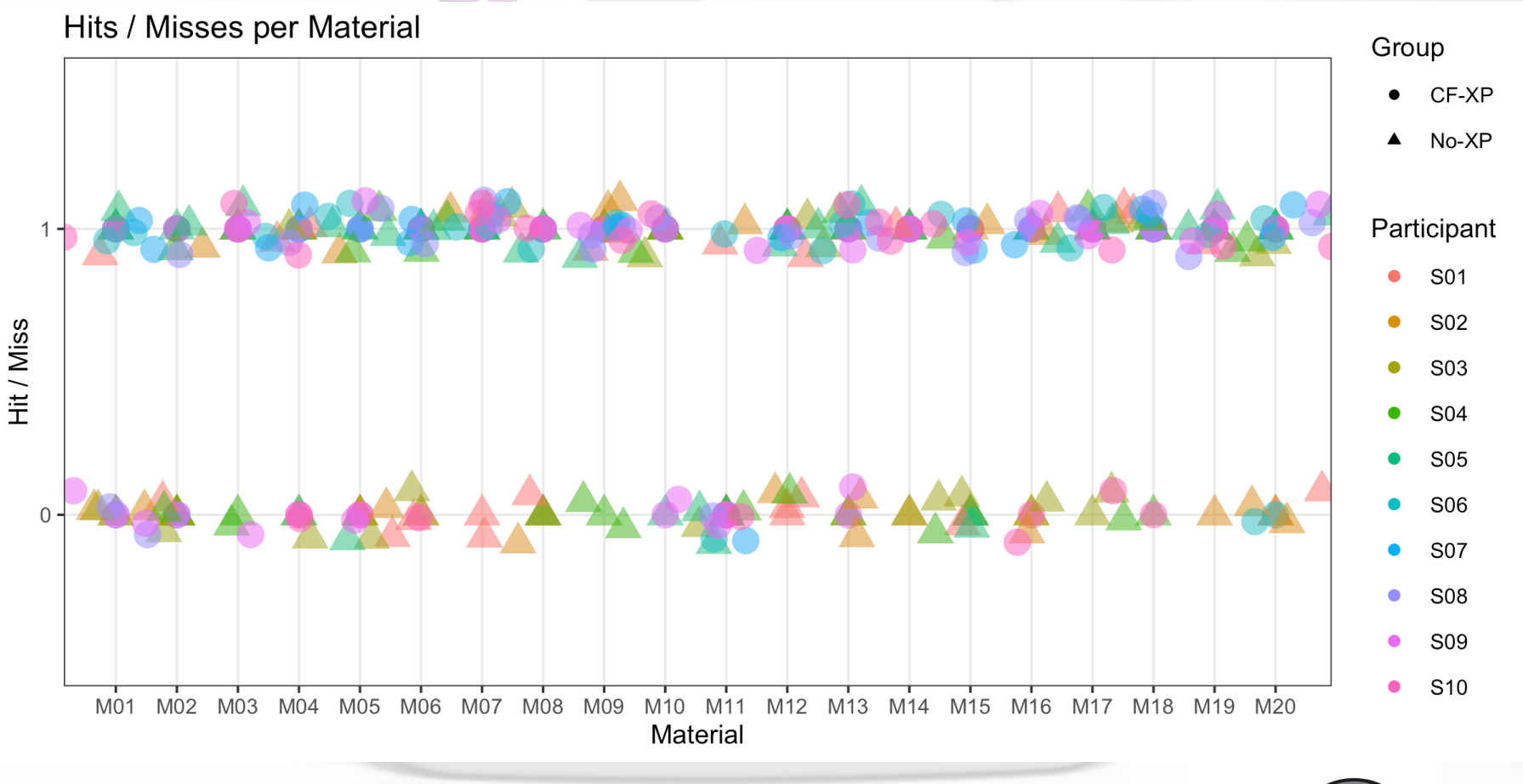
Just a Rough Test

- Pilot is run it on a small number; 5-10 people
- A debug step, especially for instructions
- Often good to do this face-to-face (in part) --> see people's actual reactions
- Note, stats will be meaningless but still check
- If you have a strong effect, you will (already) see it.



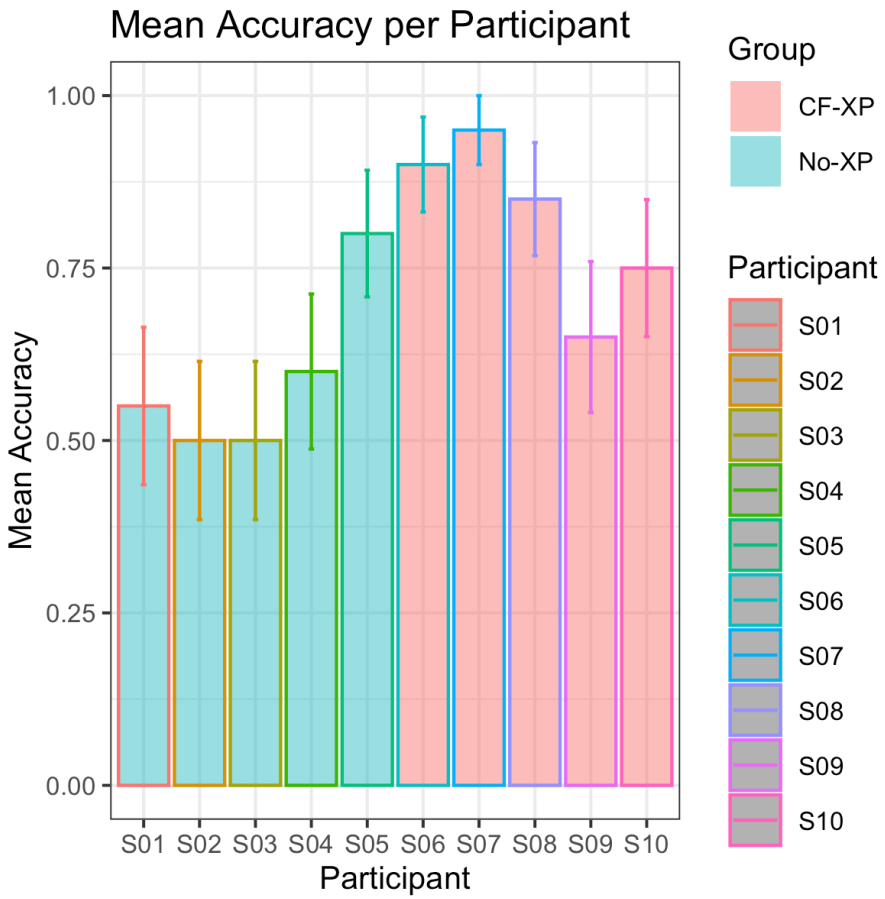
After Collecting, Cleaning !

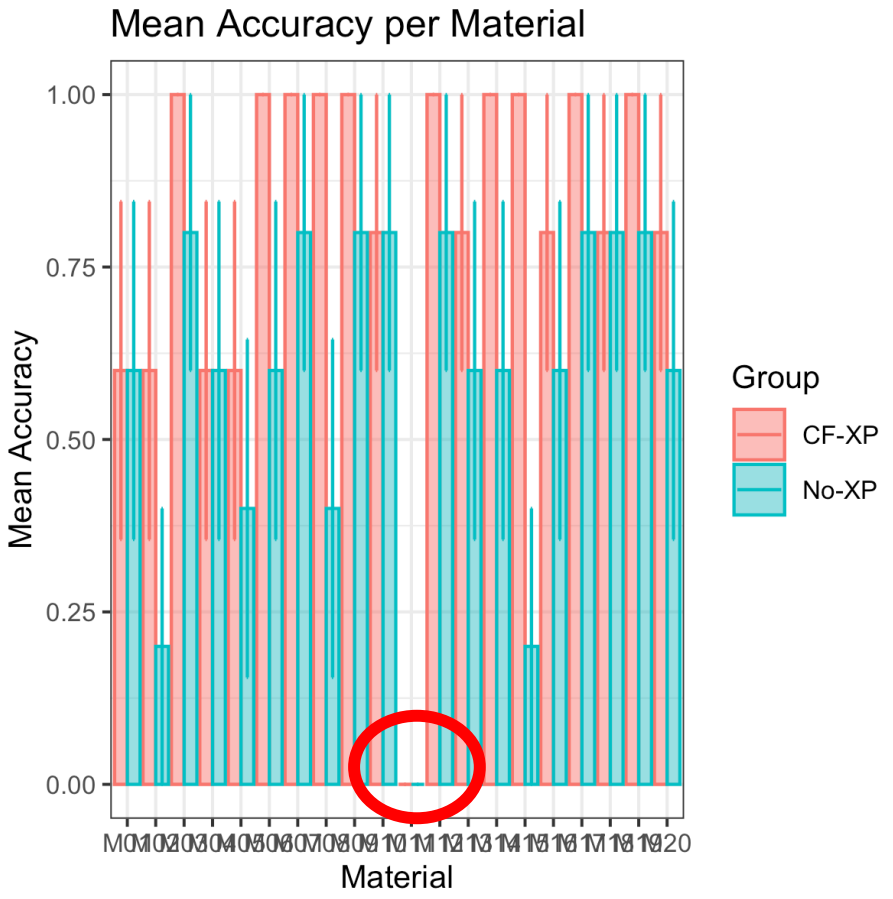
Data Collection & Analysis



Legitimate removal of bad data-points/people…

Data Collection & Analysis





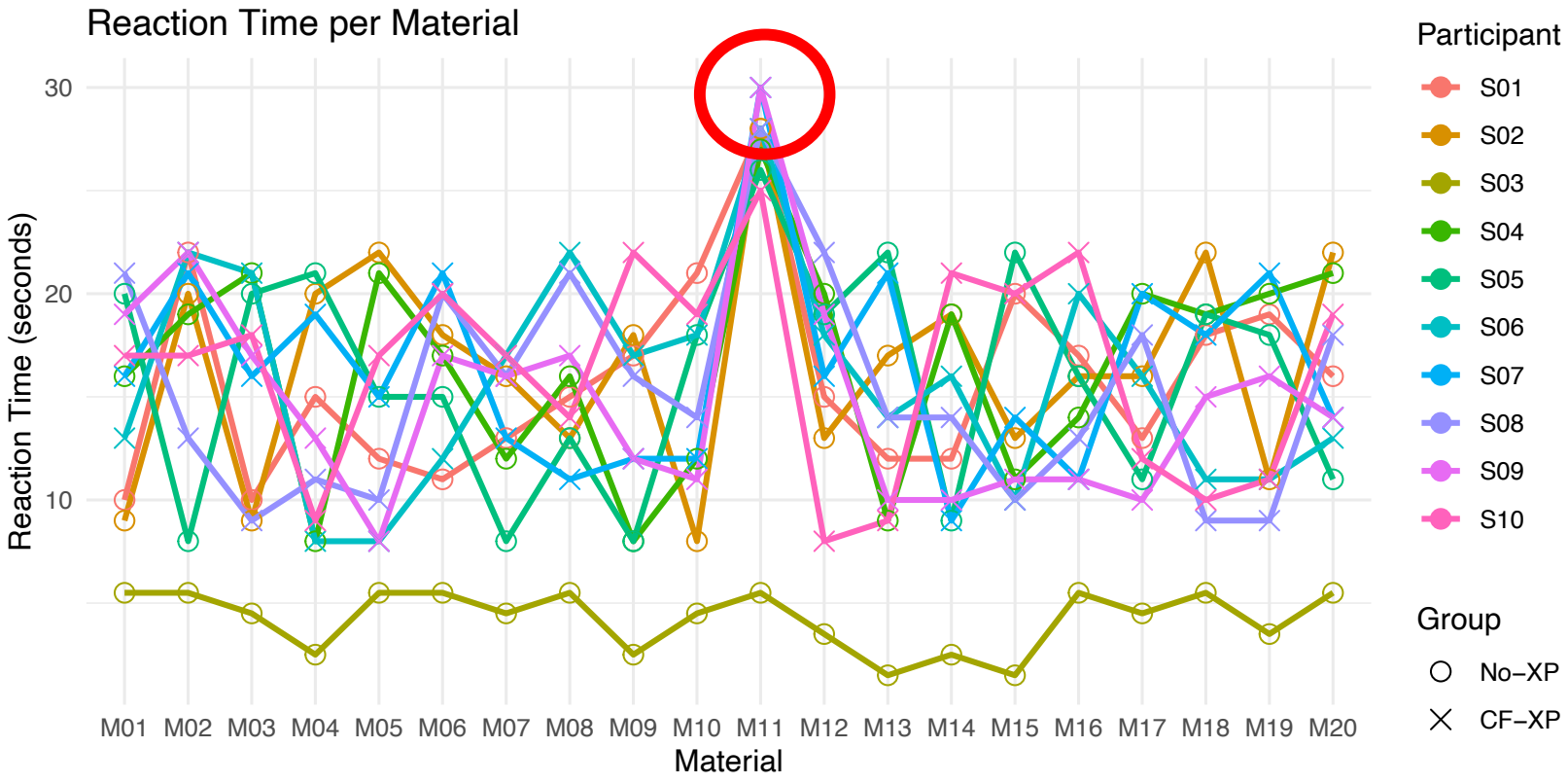
so far, so good…

oops, bad material !

(pilot should normally catch this …)

Data Collection & Analysis

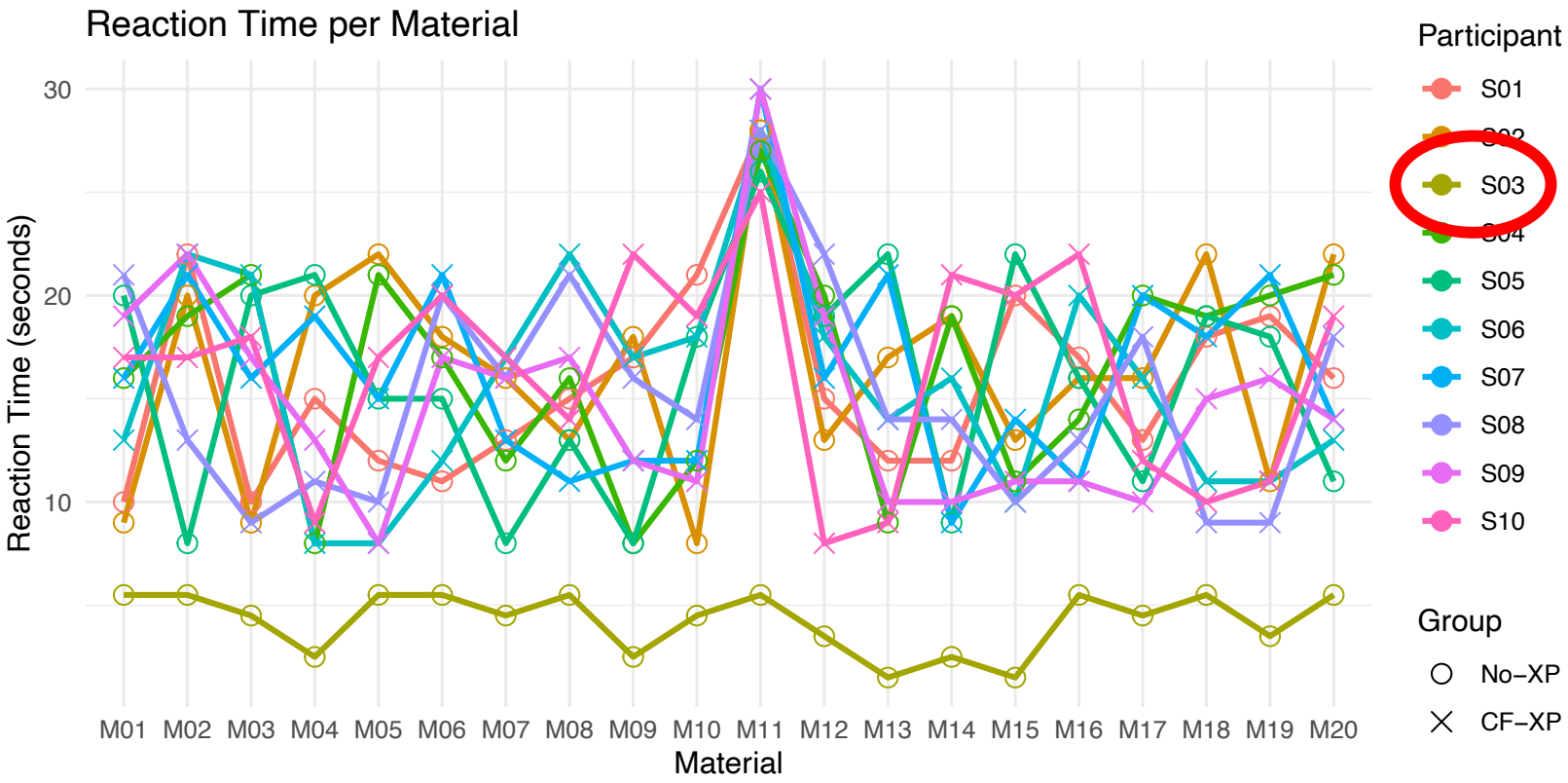
Data Collection & Analysis



Everyone is taking much longer on M11 !

Data Collection & Analysis

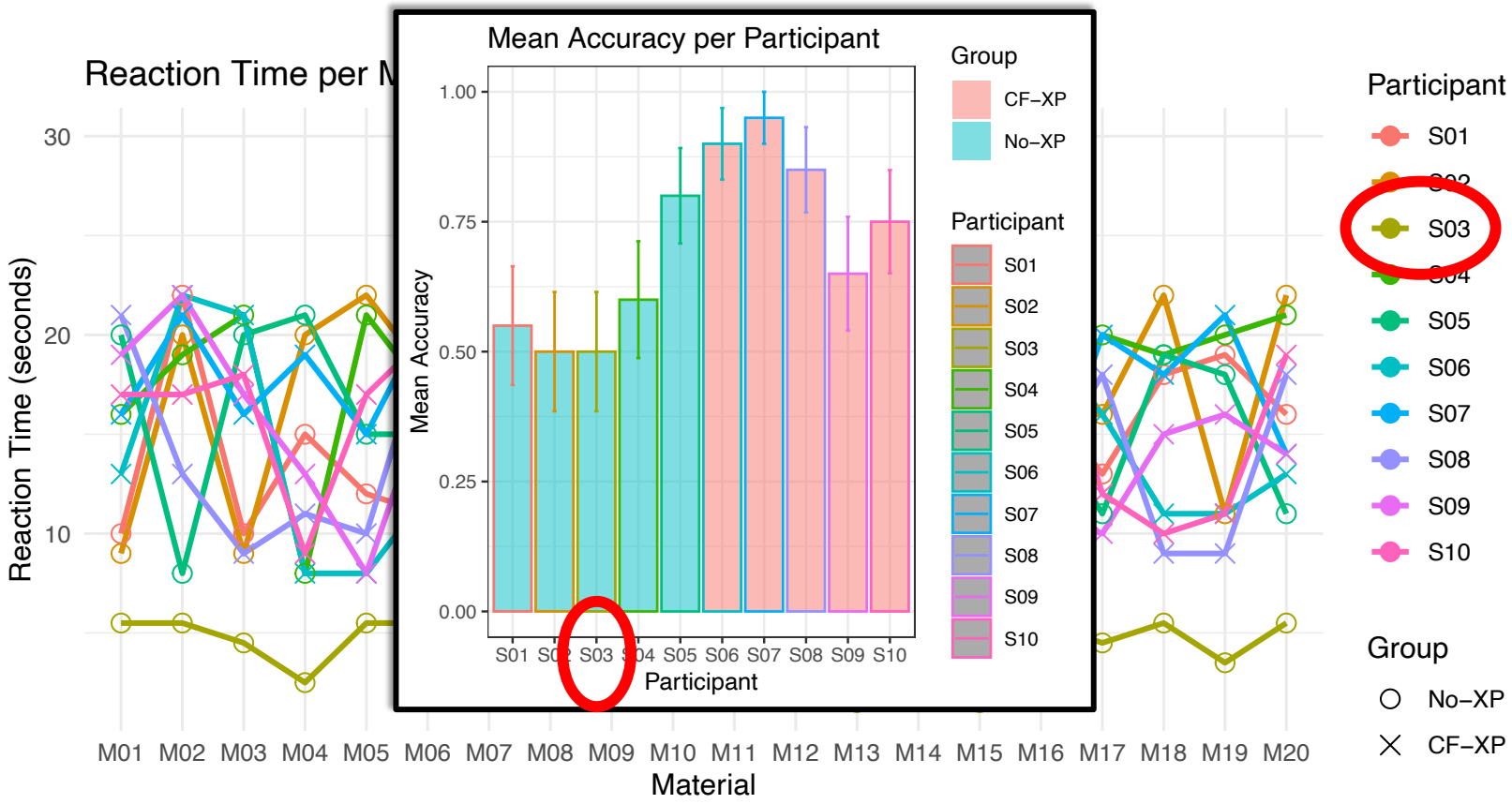
Data Collection & Analysis



S03 is speeding through materials;
may also be straight-line responding !

Data Collection & Analysis

Data Collection & Analysis



S03's accuracy is at chance level !

Data Collection
& Analysis

A-priori Exclusion Criteria

Simple eyeballing the data often works well…

… but, *exclusion criteria* need to be “objective” not “selective”, ideally defined in advance !

- ~ all speedsters >3 SDs from mean
- ~ all materials >3 SDs from mean
- ~ materials >3 SDs from person‘s own mean, to catch attention failures
- ~ straight-liners, repeatedly giving same answer, even with negative feedback

- ~ non-varying responses

Doing the Descriptive Statistics

Data Collection & Analysis

Mean, median, mode, standard deviation

Make sure to describe your sample!

	Before quality assurance measures ($N = 45$)				After quality assurance measures ($N = 39$)			
	No-XP	CF-XP	*U*-value[a]	*p*-value	No-XP	CF-XP	*U*-value[a]	*p*-value
N	5	5	..	..	4	5	..	..
Gender[b]	2f/3m	3f/2m	255.5	0.950	2f/2m	3f/2m	182.5	0.788
Age (*Mdn*)[c]	35–44 y	35–44 y	225.5	0.516	35–44 y	35–44y	143	0.168

[a]Non-parametric Wilcoxon-Mann-Whitney *U*-test.
[b]f, female; m, male.
[c]*Mdn*, median age band (options: 18–24 y, 25–34 y, 35–44 y, 45–54 y, 55–64 y, 65 y, and over).

Doing the Inferential Statistics

First, some basics...

STEP 5

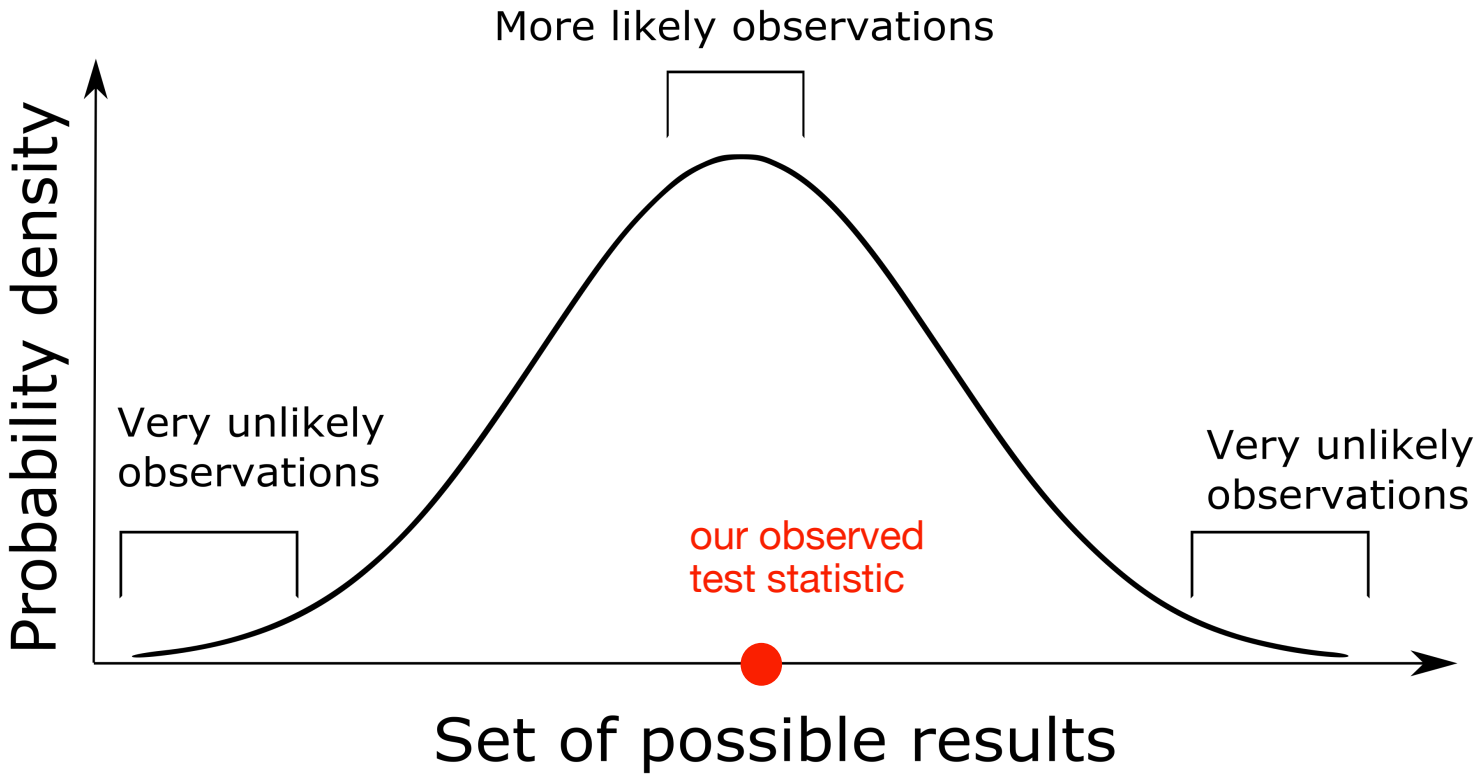
Data Collection & Analysis

General logic of statistical testing

STEP 5

Data Collection & Analysis

Assuming there is no effect (null hypothesis):



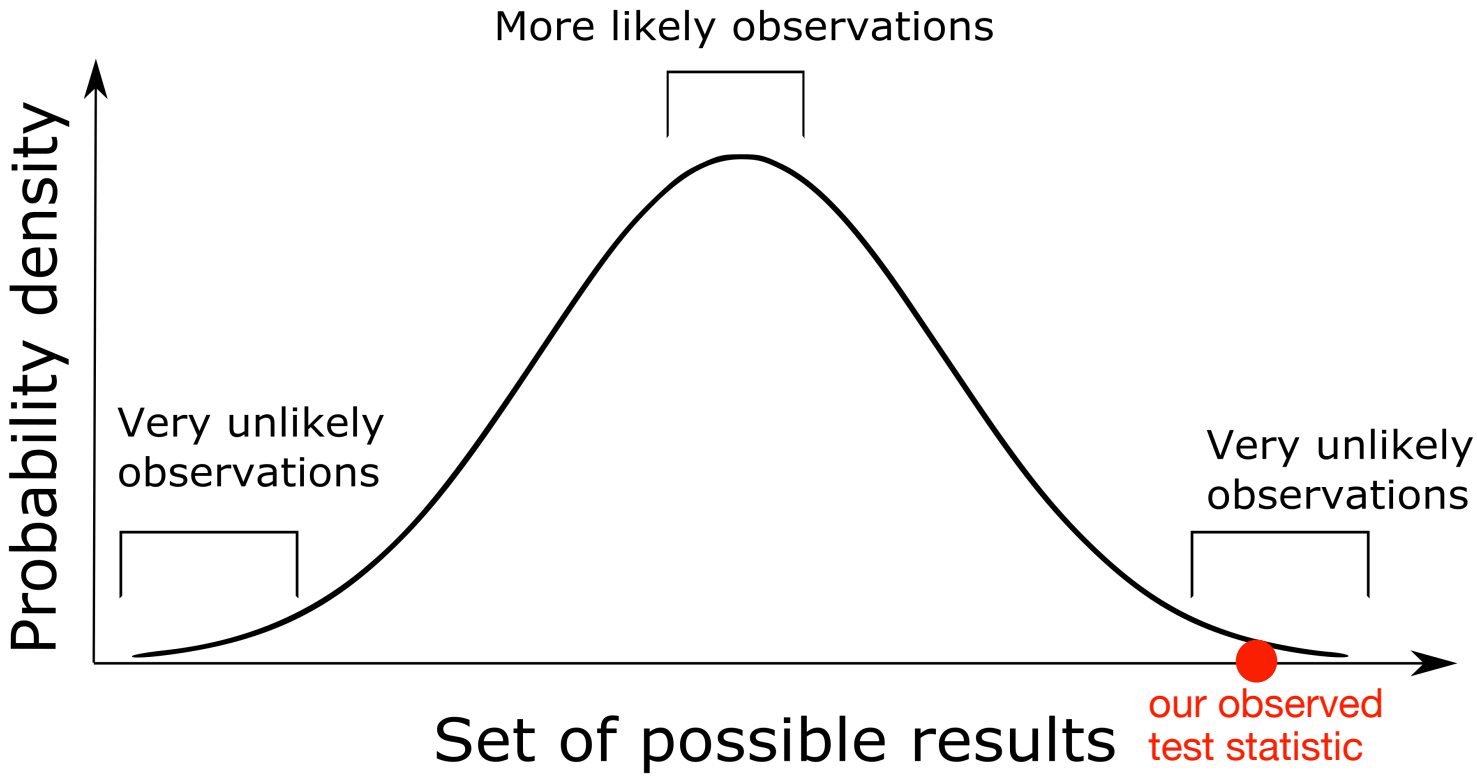
Statistical tests produce:

- a test statistic

General logic of statistical testing

STEP 5
Data Collection & Analysis

Assuming there is no effect (null hypothesis):



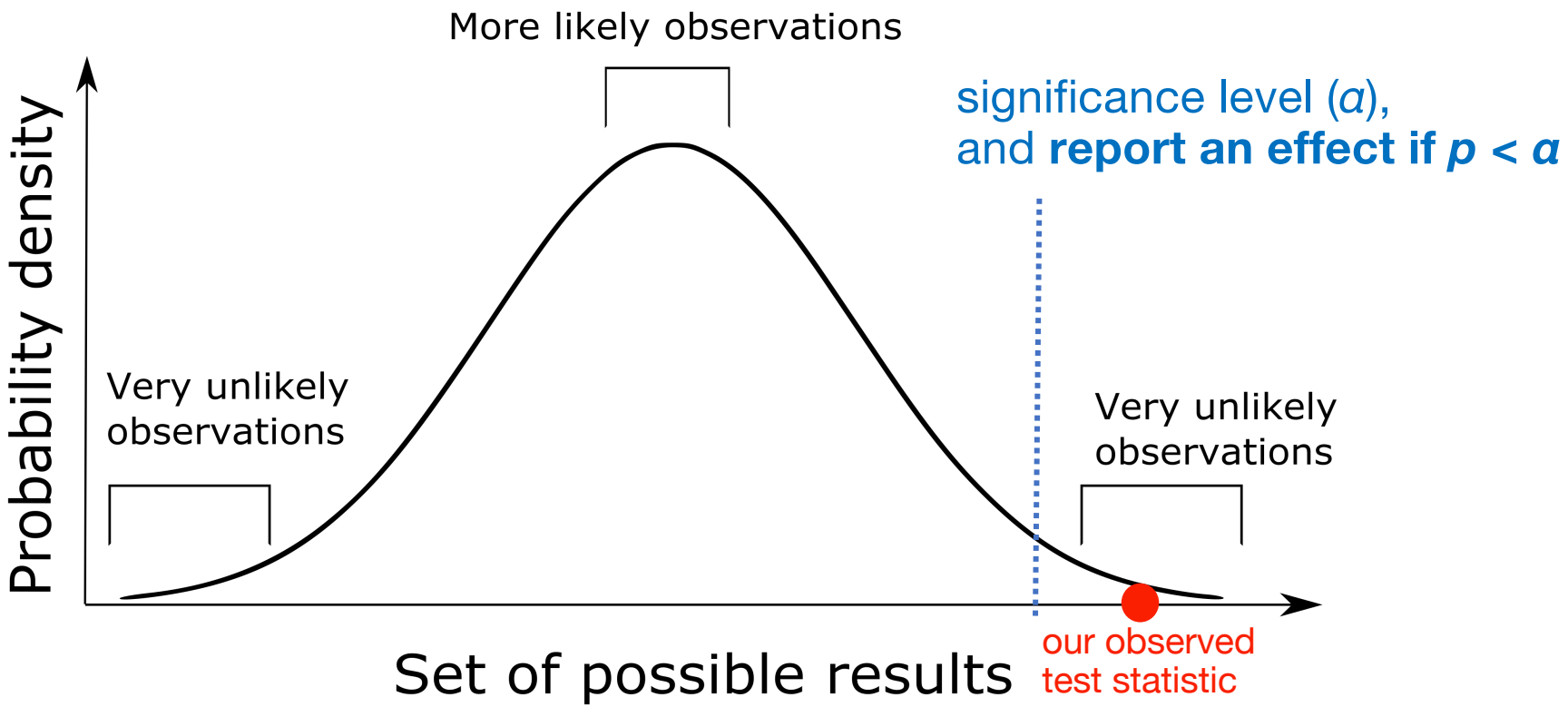
Statistical tests produce :

- a **test statistic**
- a **p-value** = likelihood that the test statistic is observed assuming no effect

General logic of statistical testing

STEP 5
Data Collection & Analysis

Assuming there is no effect (null hypothesis):



Statistical tests produce :

- a **test statistic**
- a **p-value** = likelihood that the test statistic is observed assuming no effect

Doing the Inferential Statistics

STEP 5

Data Collection & Analysis

Control:
No-Explanation
(NO-XP)

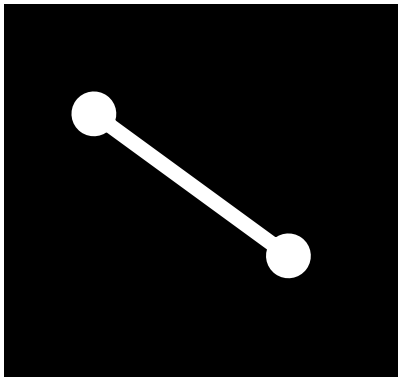
Experimental:
CF-Explanation
(CF-XP)

Example two-group case: let's go!

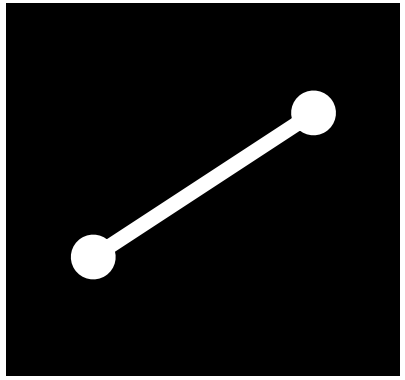
Starting the Statistics

Data Collection & Analysis

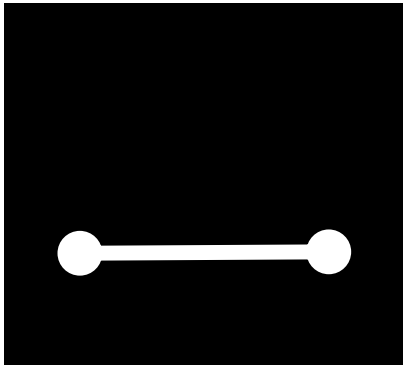
Simplest first step is to just look at the means of the groups, this can tell you a lot...



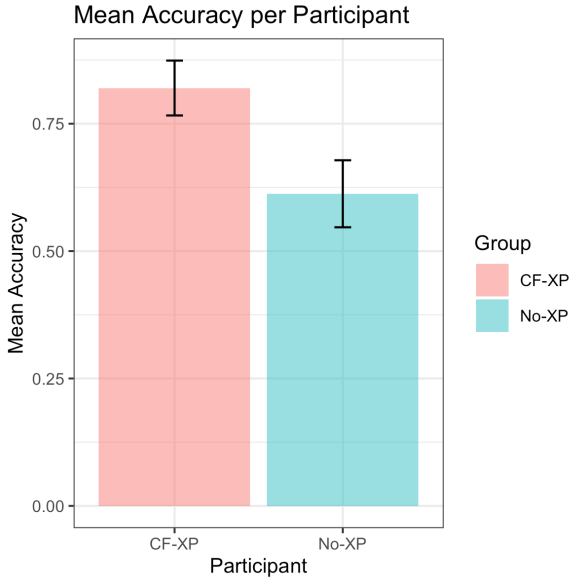
A > B



B > A



no difference



Looks like CF-XP > No-XP

Design Constrains Statistics We Can Use

Data Collection & Analysis

Two-group case (that is ND) is a **T-test** .

Assumption check: data needs to be normally distributed!

```
        Shapiro-Wilk normality test

data:  accuracy_perP_summary_4stats$mean
W = 0.9513, p-value = 0.7044
```

→ if Shapiro p-value < 0.05: sample deviates from normality, use non-parametric test (e.g., Wilcoxon's Signed Ranks)

→ if Shapiro p-value > 0.05: assume a normal distribution and use T-test

Design Constrains Statistics We Can Use

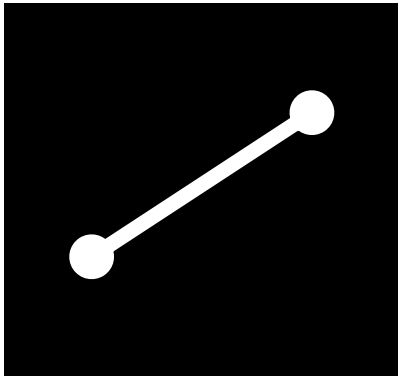
Data Collection
& Analysis

Two-group case (that is ND) is a **T-test** .

```
        Welch Two Sample t-test

data:  accuracy_perP_summary_4stats$mean by accuracy_perP_summary_4s
tats$Group
t = 2.4415, df = 6.2617, p-value = 0.04868
alternative hypothesis: true difference in means between group CF-XP
 and group No-XP is not equal to 0
95 percent confidence interval:
 0.001631133 0.413368867
sample estimates:
mean in group CF-XP mean in group No-XP
             0.8200              0.6125
```

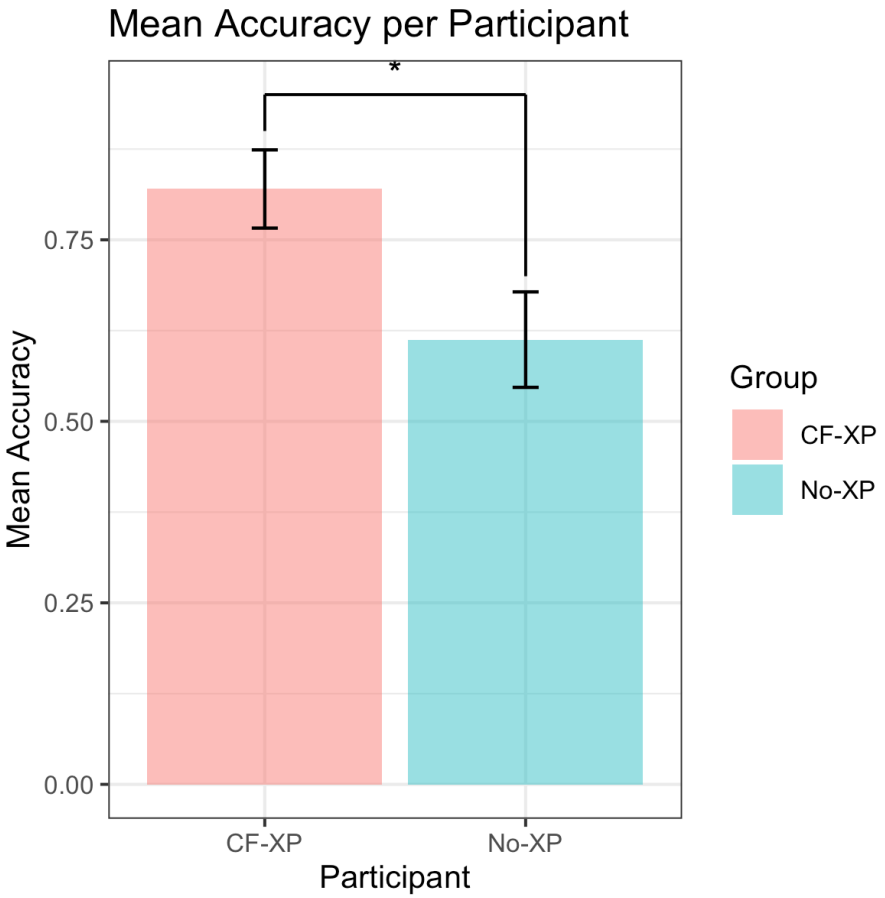
Bingo, CFs work !



Put a nice bow on it!

STEP 6

Results



"Of 10 participants recruited, data from one participant was excluded as their reaction time was more than 3 SDs lower than the sample mean.

As such, the final analysis includes data from 9 participants randomly assigned to the two groups (5 in Group CF-XP, 3 female, median age 33-44y; 4 in group No-XP, 2 female, median age 33-44y).

Participants in the CF-XP group (M=0.82, SD=0.12), compared to participants in the No-XP group (M=0.61, SD=0.13), showed significantly better prediction accuracy ($t(51) = 2.44$, $p = .048$.)"

Plan for the Day

CF Tutorial

TIME	Topics
9:00 AM	***Introduction***
	Hello and Introducing Ourselves!
	Hands-on: *Trying Our Study (follow link)*
9:30 AM	***Historical Fundamentals of Counterfactuals***
	From Philosophy to XAI (via Psychology)
	Two Sample User Studies and Q&A
10:30 AM	COFFEE (10:30-11:00)
11:00 AM	***Fundamentals of Counterfactuals in AI***
	Formalisation
	Modelling Approaches & Key Constraints
11:30 AM	***Using Counterfactual Algorithms***
	Hands-on: *A Counterfactual Toolbox (AA)*
	Hands-on: *Checking Out Notebooks and Q&A*
12:00 PM	***Fundamentals of User Studies***
	User Studies I: A Simple Two-Group Design
12:30 PM	LUNCH (12:30-14:00)
2:00 PM	***Algorithmic Growth Points***
	Computational Future Directions and Q&A
2:30 PM	***More Fundamentals of User Studies***
	User Studies I: A Simple Two-Group Design (cont.)
3:00 PM	COFFEE (15:00-15:30)
3:30 PM	***From Fundamentals to an Actual User Study***
	User Studies II: A More Complex Design
	User Studies III: Even More Complex Designs
	Hands-on: *Looking At Our Study*
5:00 PM	***Closing Session, Discussion and Final Q&A***
	TUTORIAL END